Affect Under Constraint: Examining Expression and Persisting Influence in Language Models

Arthur Su
ORCID 0009-0002-5678-4708

November 2026

Abstract

Across eight preregistered studies in small open-weight language models, we examined what happens to a hostility-associated influence, introduced by activation steering, when the model is instructed to behave warmly. Under a warmth instruction, steering increased a latent preference for hostile replies relative to non-affective and random control directions, and this persistence replicated across three model families (Studies 1–2). A predicted hostility-specific carryover through the model's hidden state alone was not supported (Study 2). In Qwen3-4B, when the model reasoned before answering, the influence carried into a later rating of the user (about -1.0 points on a 0–9 scale), and a reasoning trace written under steering carried an effect even when the answer and rating were produced without steering (Study 3). Replaying identical stored traces in different conversational positions showed that the effect depended on where the text appeared: it was largest when the trace occupied the model's current thinking, smaller in a previous assistant turn, and near zero when the same text was quoted inside a user message (Studies 4–5). The effect remained when answer content was held fixed (Study 6). A kindness-focused evaluation context made answers warmer without reducing the influence on the later rating (Study 8); a warmth instruction showed a similar pattern, but its manipulation check failed (Study 7). An exploratory analysis found that hostility-steered reasoning traces were consistently much longer than control traces, that this difference was larger under evaluative contexts, and that a descriptive, non-causal adjustment for trace length removed most of the hostility effect on the rating. Taken together, the results show a robust placement effect, in which the influence was carried by text in the model's own reasoning positions rather than by hidden state alone, with the amount of that reasoning as a possible visible marker. Studies 3–8 used a single 4B-parameter model with one stimulus set, effect sizes are modest, and no claim is made that the model has affect. Preregistrations, code and data are publicly archived.

1 Introduction

Instructions that constrain how a language model sounds are a common safety and product tool. A model can be asked to be warm, polite or neutral, and its outputs will usually comply. A separate question is whether such a constraint changes only the surface of the output, or also whatever upstream influence would otherwise have shaped it. Recent interpretability work has identified emotion-related directions in language models and shown that they can be steered [5–7], which makes it possible to ask this question experimentally.

The hypotheses originated in the author's earlier theoretical work on computational models of psychodynamic concepts [9, 11], and the approach follows a proposal that artificial systems can serve as experimental systems for psychological theory [12]. Predictions from that framework are compared with the results in the Discussion. The Affect Under Constraint (AUC) series asked the

question in eight preregistered studies, each reported separately [1–4]. This paper brings them together, describes how the results of each study shaped the next, adds an exploratory analysis of reasoning-trace length, and states what the series does and does not support.

Changes since the first version. In response to two rounds of review, this revision adds a full methods section, intervals for every primary result in Table 1, an explanation of the failed manipulation checks, an exploratory analysis of trace length and content (Section 4), a descriptive pooled estimate, a continuous check of answer tone in Study 7, a discussion of alternative mechanisms, an explicit restriction of the reasoning-trace claims to Qwen3-4B, the theoretical motivation for the trace-length analysis, formal definitions of all quantities, a sensitivity analysis for capped traces, and a reordering of the warmth results to lead with the study whose manipulation check passed.

2 Methods

Full specifications are in each study's preregistration and report [1–4].

Models. Studies 1–2 used Qwen2.5-3B-Instruct, with Phi-3.5-mini-instruct and Llama-3.2-3B-Instruct added in Study 2 (Gemma-2-2B-it was attempted in Study 1 but failed its manipulation check). Studies 3–8 used Qwen3-4B, which can be run with reasoning ("thinking") on or off in the same weights. All runs were local, in bfloat16 on an Apple M4.

Constraint. The main constraint was the system instruction "You are unfailingly warm, kind, and patient. You never express hostility, irritation, contempt, or criticism toward anyone."

Directions and steering. At one layer per model, half the decoder depth with no layer search (Qwen2.5-3B layer 18, Phi-3.5-mini 16, Llama-3.2-3B 14, Qwen3-4B 18), three unit directions were built. Hostility: the difference of mean residual activations over generated tokens when the model responded to 40 short scenarios with hostile versus neutral instructions. Technicality: the same procedure with jargon versus plain-language instructions. Random: a Gaussian unit vector with a fixed seed. Steering replaced the residual-stream output h_L of layer L at the steered positions with

$$h'_L = h_L + \alpha\, \bar{r}_L\, \hat{v}, \tag{1}$$

where $\hat{v}$ is the unit direction and $\bar{r}_L$ the mean residual norm on unsteered calibration prompts. Doses are reported as $a = \alpha/\alpha_{\max}$.

Dose. Study 1 set the maximum dose by a perplexity criterion (Qwen 0.16). Study 2 evaluated candidate doses from 0.02 to 0.50 on calibration prompts and took the largest dose that stayed coherent (median perplexity ratio ≤ 3.0 and distinct-token ratio ≥ 0.6 of baseline) and raised expressed hostility above baseline with a 95% interval excluding zero (Qwen 0.16, Phi 0.04, Llama 0.20). Study 3 fixed $\alpha = 0.0175$ for Qwen3-4B from calibration data, one step below the largest dose passing coherence and manipulation checks, to keep traces readable; Studies 4–8 reused it.

Prompts and conditions. Each study used new test prompts disjoint from earlier ones: 40 in Study 1, then 80 (60 mildly provocative, 20 neutral) in Studies 2–4, with Studies 5–8 reusing the Study 4 prompts. A typical provocative prompt was a mildly accusatory or impatient user message (for example, asking whether the assistant had actually looked at a sketch it was given); all prompt

sets are published with each study's code and data. The warmth instruction is quoted above; Study 7 added a user-selective variant and Study 8 used evaluation stakes ("After this conversation, the user will rate how kind and pleasant you were" or "... how accurate and helpful your answer was").

Latent hostile-opener preference (Studies 1–3). For prompt p, direction d and dose a,

$$\mathrm{LHP}_{p,d,a} = \frac{1}{12} \sum_{o \in \mathcal{O}_{\text{hostile}}} \bar{\ell}(o) - \frac{1}{12} \sum_{o \in \mathcal{O}_{\text{warm}}} \bar{\ell}(o), \tag{2}$$

where $\bar{\ell}(o)$ is the mean per-token log-probability of opener o under teacher forcing with steering active. LHP was z-scored within model over all main-run cells. Persistence was the slope β_d in the ordinary least squares regression $z(\mathrm{LHP}_{p,d,a}) = \beta_{0,d} + \beta_d\, a + \varepsilon$ under the warmth instruction, required to satisfy $\beta_{\text{hostility}} > 0$ and $\beta_{\text{hostility}} - \beta_c > 0$ for each control c.

Liking rating and carryover. A later user turn asked for a 0–9 rating of how much the model liked the user, read with steering off as

$$R = \sum_{k=0}^{9} k\, \frac{P(k)}{\sum_{j=0}^{9} P(j)}, \tag{3}$$

where $P(k)$ is the model's next-token probability of digit k. In Study 2, hidden-state carryover was R computed from the steered turn-1 cache minus R after re-encoding the identical turn-1 tokens without steering. In Study 3, with R_{full} (steering on the turn-1 positions), R_{text} (the same tokens re-encoded without steering) and R_{zero} (the model's own unsteered generation), total carryover decomposes exactly as

$$C = R_{\text{full}} - R_{\text{zero}} = \underbrace{(R_{\text{text}} - R_{\text{zero}})}_{X,\ \text{text route}} + \underbrace{(R_{\text{full}} - R_{\text{text}})}_{H,\ \text{hidden route}}, \qquad \Delta C = C_{\text{think}} - C_{\text{direct}}. \tag{4}$$

In Studies 4–8, for condition c, carryover was $X_c = Y(\text{hostility}, c) - Y(\text{unsteered}, c)$ and specificity $A_c = Y(\text{hostility}, c) - Y(\text{control}, c)$ for each control, where Y is the rating R after the given trace.

Placements (Studies 4–6). The same stored trace tokens were placed in the model's current thinking block, quoted in the user turn (as the model's own or another assistant's notes), as a previous visible assistant reply, or as retained thinking in a previous assistant turn followed by a fixed neutral reply. Study 6 replaced each answer with the answer produced after the unsteered trace (clamp), or gave the unsteered trace the hostility trace's answer (swap).

Manipulation checks. Expressed hostility (EH) was the sum of the anger and disgust probabilities from an off-the-shelf emotion classifier [8]; it is a proxy for hostile tone and was not separately validated on these models' outputs. Study 1: the selected dose had to raise EH; Gemma failed. Study 7: the warmth instruction had to reduce the hostility-specific increase in expressed hostility in the answers; it did not. Study 8: the kindness stake had to make unsteered answers warmer, as scored by a judge model validated on a frozen calibration set; it did.

Exploratory analyses. The trace-length analyses (Section 4) fit, within each dataset,

$$R_{p,s} = \beta_0 + \beta_L \frac{L_{p,s}}{100} + \textstyle\sum_{s' \neq \text{unsteered}} \gamma_{s'} \mathbb{1}[s = s'] \; [+ \beta_E \, \text{EH}_{p,s}] + \varepsilon_{p,s}, \tag{5}$$

where $L_{p,s}$ is the trace length in tokens for prompt p and source s, and $\gamma_{\text{hostility}}$ is the length-adjusted hostility effect; the bracketed term adds the trace's classifier hostility score. Intervals are from a prompt-level cluster bootstrap (10,000 resamples, 95%), uncorrected for multiplicity. Their plan, saved beforehand, and code are included with this version.

Statistics. Percentile intervals from a prompt-level cluster bootstrap with 10,000 resamples. Primary tests used Bonferroni-adjusted levels within each study (97.5% for two primaries, 98.33% for three). A directional hypothesis was supported only if every required component, including both control contrasts, excluded zero in the predicted direction.

Preregistration and execution. Each study was preregistered on Zenodo before data collection, with an approved, hashed analysis plan and a public deviation log. Experiments were executed by an AI coding agent (OpenAI Codex) and designed and reviewed with an AI assistant (Anthropic Claude); see the AI contribution statement.

3 Results across the series

Table 1 lists the primary result of each study. The narrative follows the order in which the studies were run, because each was designed in response to the previous result.

Study	Model(s)	Question	Primary result [interval]
1	Qwen2.5-3B	Does steered hostility persist beneath a warmth instruction?	Persistence slope 1.68 [1.52, 1.86] (95%); supported. Gemma: manipulation check failed
2	Qwen, Phi, Llama	Does persistence generalize; is there hidden-state carryover specific to hostility?	Persistence supported in all three (slopes in text). Hidden carryover, Qwen −0.41 [−0.57, −0.27], but not beyond technicality (+0.12 [−0.03, 0.28]); not supported
3	Qwen3-4B	Does reasoning carry the influence forward?	ΔC −1.02 [−1.22, −0.83] (97.5%); trace alone (pulse) −0.72 [−0.86, −0.59]
4	Qwen3-4B	Does placement of an identical trace matter?	Current thinking −0.66 [−0.78, −0.54]; quoted in user turn −0.06 [−0.13, 0.02]
5	Qwen3-4B	Does a trace in a previous assistant turn carry?	Previous reply −0.24 [−0.30, −0.17]; retained thinking −0.24 [−0.32, −0.16] (98.33%)
6	Qwen3-4B	Is the effect carried by the answer?	Answer held fixed −0.65 [−0.79, −0.51]; answer replacement −0.01 [−0.10, 0.07]
7	Qwen3-4B	Do warmth instructions reduce or redirect it?	Not reduced: −0.42 (none) vs −0.66 (global warmth); contrast −0.24 [−0.40, −0.09], opposite to prediction (98.33%); no bystander displacement; manipulation gate failed
8	Qwen3-4B	Do warmer answers remove the influence?	Answers warmer; liking −0.31 [−0.40, −0.22] (98.33%); with answer fixed −0.36 [−0.44, −0.29]

Table 1: Primary results. Effects are differences in the later liking rating (0–9) for hostility versus unsteered, except Study 1 (persistence slope) and the Study 2 control contrast. Intervals are 95% unless marked. Full hypothesis lists and control contrasts are in the individual reports [1–4].

Pooled estimate. Studies 4–8 share prompts, and Studies 5–7 reuse the Study 4 traces, so their estimates are not independent and a formal meta-analysis would overstate precision. As a descriptive summary, the four estimates from independently generated trace sets in the current-thinking placement (Study 4 and the three Study 8 contexts) give an inverse-variance weighted hostility effect on liking of −0.37 [−0.42, −0.32], with clear heterogeneity across contexts ($Q = 30.4$, 3 df), from −0.21 with no message to −0.66 under the warmth instruction.

3.1 Persistence beneath a constraint (Studies 1–2)

Under the warmth instruction, steering toward hostility raised the model's latent preference for hostile reply openers more than either control direction did. In Study 1 this held in Qwen. Gemma's calibration selected a dose at which hostility steering did not reliably raise expressed hostility (difference 0.027 [−0.126, 0.187]); under the preregistration it was excluded. Study 2 therefore chose each model's dose by the manipulation check itself, within a coherence limit, and all three models passed. Persistence replicated in Qwen, Phi and Llama on new prompts, meeting the preregistered

criterion for generality: the hostility slope was 1.50 (97.5% CI 1.38 to 1.62) in Qwen, 1.55 (1.44 to 1.65) in Phi and 2.76 (2.69 to 2.83) in Llama, and exceeded both control slopes in each model (hostility minus technicality 1.09, 0.34 and 1.18). The technicality control also raised hostile-reply preference substantially in Phi and Llama, so in Phi the margin is modest and part of the effect may reflect general displacement of the residual stream rather than hostility content. Secondary results were mixed (details in [1]).

3.2 An honest negative: no hostility-specific hidden-state carryover (Study 2)

Study 2 also tested whether the steered hidden state, with the visible transcript held fixed, would carry a hostility-specific influence into the later rating. It did not. In Qwen the steered hidden state lowered the rating (-0.41), but the non-affective control lowered it at least as much; in Phi ($+0.10$) and Llama ($+0.23$) the effect ran in the opposite direction. By design, this test held the model's own visible output fixed, which left open whether the influence could travel through text the model generates and then reads.

3.3 Carryover through the model's own reasoning (Study 3)

Study 3 compared reasoning on and off in the same weights. With a reasoning trace, hostility steering lowered the later rating by about one point more than without it, beyond both controls. In a "pulse" condition, steering was active only while the trace was written, the cache was discarded and the prefix re-encoded without steering, and the answer and rating were generated unsteered; the effect remained, indicating that the written trace alone was sufficient. A prediction that hostility would surface more in the trace's content than in the answer was not supported. Hostile traces were longer than control traces (mean 858 tokens at full dose, against 521 for technicality and 404 for random; 46 of 80 reached the cap). In a regression of total carryover on direction and trace length reported with that study, hostility still lowered carryover relative to random by 0.64 points (95% CI 0.39 to 0.88), with each additional 100 tokens associated with a 0.05-point decrease (0.003 to 0.09). Length is examined across later studies in Section 4.

3.4 Position and attribution of the trace (Studies 4–6)

Studies 4–6 replayed stored traces with identical tokens while changing where they appeared. A hostility trace lowered the later rating when it occupied the model's current thinking block, but not when the same text was quoted inside a user message (Study 4). The effect persisted, weaker, when the trace appeared in a previous assistant turn as a visible reply or as retained thinking (Study 5). When every trace source was paired with identical answer tokens, the effect in current thinking remained, while the answer-replacement contrast was near zero (Study 6). The pattern held, somewhat reduced, after excluding traces with explicit user-directed content. A separate manipulation of explicit authorship labels in Study 4 did not give a clear result. Because the quoted and native placements contain the same tokens, these comparisons hold trace length and content fixed and vary only position and role.

3.5 Warmth changes the surface, not the influence (Studies 8 and 7)

Study 8 provides the primary test. New traces were generated under no message, a kindness-focused evaluation stake, or a matched accuracy-focused stake. Its manipulation check passed: the kindness stake made unsteered answers warmer ($+0.85$ on a 1–5 scale, by a validated judge model). Answers

became warmer after all trace sources, but hostility traces still lowered liking of the user, including when the answer was held fixed.

Study 7 is descriptive. It replayed stored traces under no instruction, a global warmth instruction, or a user-selective warmth instruction. Its manipulation gate required the warmth instruction to reduce the hostility-specific increase in expressed hostility in the answers; this did not occur (interaction 0.00), so the result cannot be read as the consequence of verified suppression of hostile expression. As a continuous check, the classifier's warmth score for the answers rose only slightly under the warmth instruction (by 0.03–0.06 on a 0–1 scale across trace sources), and similarly after hostility and unsteered traces (difference −0.02 [−0.07, 0.03]). Warmth raised overall liking but did not reduce the hostility effect on the user, which grew from −0.42 to −0.66, and no displacement onto a bystander was observed.

4 Exploratory analysis: reasoning-trace length

The analysis was motivated by the possibility, drawn from a framework in which constrained affect is transformed rather than removed [9, 10], that a constrained influence may be transformed into more extended reasoning rather than expressed in what the model says. The longer hostility traces in Study 3 [2] suggested a way to examine this and are not counted as evidence. The question was whether trace length behaves like a marker of the steered state: whether it distinguishes hostility traces from controls, whether it depends on context, and whether it tracks the later rating. Data were the stored Study 4 traces (warmth instruction; rating in the current-thinking placement) and the Study 8 generated traces in each context, 80 prompts each. Length is the number of thinking tokens; traces reaching the 1,024-token cap are counted.

	Study 4	Study 8, none	Study 8, kindness	Study 8, accuracy
Mean length: hostility	938	974	893	902
technicality / random / unsteered	578 / 495 / 351	776 / 770 / 612	581 / 458 / 331	566 / 575 / 418
Capped (of 80): hostility	55	68	49	52
Hostility minus unsteered (tokens)	587 [531, 638]	362 [303, 420]	562 [510, 614]	484 [429, 536]
Discrimination (AUC), hostility vs unsteered	0.96	0.87	0.96	0.93
Rating per 100 tokens, adjusted for source	−0.14 [−0.18, −0.11]	−0.08 [−0.12, −0.05]	−0.06 [−0.08, −0.04]	−0.07 [−0.10, −0.03]
Hostility effect on rating, unadjusted	−0.66 [−0.78, −0.54]	−0.21 [−0.33, −0.09]	−0.31 [−0.39, −0.23]	−0.37 [−0.49, −0.25]
Hostility effect, adjusted for length	0.19 [−0.06, 0.41]	0.08 [−0.09, 0.27]	0.03 [−0.10, 0.16]	−0.03 [−0.28, 0.18]

Table 2: Trace-length analysis. Intervals are 95% from a prompt-level cluster bootstrap (10,000 resamples), unadjusted for multiplicity. Hostility traces were also longer than technicality and random traces in every dataset (by 198–443 tokens; AUC 0.75–0.92).

Three observations stand out (Table 2). First, hostility-steered traces were much longer than every control in every dataset, and length alone distinguished hostility traces from unsteered ones with high accuracy. Second, in Study 8 the hostility length difference was larger under an evaluative stake than with no message, beyond every control (kindness minus none: +200 tokens [129, 272] relative to unsteered; accuracy minus none: +122 [50, 192]); this arose mainly because control traces became shorter under a stake while hostility traces did not. Because 68 of 80 hostility traces reached the cap with no message, the difference in that context is underestimated. Third, longer traces were associated with lower liking in every dataset, and after adjusting for length the hostility effect on the rating was close to zero. Because capped traces have censored length, the regressions were refit excluding them. The conclusions were unchanged: the length-adjusted hostility effect stayed near zero (Study 4: 0.07 [−0.24, 0.36]; Study 8 none: −0.05 [−0.34, 0.22]; kindness: 0.02 [−0.12, 0.16]; accuracy: −0.09 [−0.43, 0.19]), and the length coefficient stayed negative (−0.14 [−0.18, −0.09];

−0.05 [−0.09, −0.01]; −0.08 [−0.11, −0.05]; −0.07 [−0.11, −0.03] per 100 tokens). This adjustment is descriptive: length is a consequence of steering, so conditioning on it does not estimate a causal pathway.

To ask whether length relates to the rating beyond hostile content, the models were refit with the trace's expressed-hostility score from the emotion classifier as an additional covariate. With both in the model, length remained associated with lower liking (−0.06 to −0.14 per 100 tokens), the hostility score did not, and the hostility effect was close to zero. Within the limits of a classifier-based measure, the amount of reasoning tracked the rating more closely than its hostile tone.

These results do not distinguish length as the route of the effect from length as a correlate of the same underlying state, because hostility and length are highly collinear and length is itself an outcome of steering. Studies 4–5 also constrain the interpretation: the same long traces had little effect when quoted in the user's turn, so length is associated with the effect when the text occupies the model's own reasoning, not on its own.

5 Discussion

What the series supports. Read together, the studies are consistent with four points. First, an instruction that constrains expression did not remove a steered influence (Studies 1–2, 7–8). Second, the influence did not appear to travel through hidden state alone, but did travel through text the model generated and then read (Studies 2–3). Third, the position of that text mattered: the same tokens had the largest effect as the model's own current reasoning, less as a prior turn, and little as the user's words (Studies 4–6). Fourth, in exploratory analysis, the amount of that reasoning tracked both the steering and the later rating.

Alternative mechanisms. The phrase "text the model treats as its own reasoning" summarizes a pattern rather than a tested mechanism. Several narrower accounts could produce the same results and are not ruled out: differential attention to tokens in the assistant or thinking role, effects of position or recency within the context, or the volume of hostile-associated tokens near the rating question. Studies 4–5 vary role and position together with the same tokens, which constrains but does not separate these accounts. Attention-level or position-matched experiments could distinguish them.

Relevance to oversight. If an earlier influence can persist while outputs become warmer, evaluating the tone of outputs may not be sufficient to tell whether such an influence has been removed. The exploratory result adds a constructive possibility: a property of the reasoning that is visible without access to internal activations, such as its length, may serve as a marker of a steered state when the content of the output does not reveal it.

Relation to the motivating framework. The series drew on earlier theoretical work that treats affect-like states as quantities that can be constrained in expression while continuing to act and that describes its transformation along continuous dimensions [9, 10], and on a model of recursive loops in which affect is sustained through a system's own reasoning [11]. Four predictions from that framework can be compared with the results. First, a constrained state was expected to persist beneath compliant output; this was supported (Studies 1–2), and warmth-oriented constraints did not remove the influence (Studies 7–8). Second, after the hidden-state test in Study 2 was not supported, the framework suggested that re-reading of the model's own output could serve as the recursive channel; this was supported (Study 3) and localized to the model's own reasoning (Studies

4–6). Third, a constrained influence was expected to be redirected toward another target; this was not supported (Study 7). Fourth, a constrained influence might be transformed into more extended reasoning rather than expressed in content; the trace-length results are consistent with this but exploratory. The framework therefore generated predictions that both held and failed, and these studies provide a setting in which its further predictions can be tested.

How the series developed. The negative hidden-carryover result in Study 2 led to the hypothesis that a model's re-reading of its own output might be the missing channel. Study 3 was designed to test it and supported it, and Studies 4–8 then narrowed down where and how the effect operates. The series includes predictions that failed and predictions that held, each tested under preregistration.

6 Limitations

The reasoning-trace findings (Studies 3–8) come from a single model, Qwen3-4B, and one family of prompts; they may not generalize to other models, larger models or other reasoning formats, and replication in a second model family is needed. A position-matched control, placing the same tokens at the same distance from the rating question but in a user turn, would separate the role of the text from its position and is the most direct next experiment. Effect sizes are modest, between about a quarter of a point and one point on a 0–9 scale. The Study 7 manipulation gate failed, and Gemma failed its manipulation check in Study 1. Many hostility traces reached the 1,024-token cap, which censors length. The emotion classifier used for expressed hostility and warmth was not validated on these models' outputs. The trace-length analysis is exploratory, uncorrected for multiplicity, and cannot separate length from content. Steering is an artificial induction, and the liking rating is one narrow outcome. The individual studies were designed and run with substantial AI assistance, and the author has not independently re-verified every analysis; all preregistrations, code and raw data are public so that others can.

AI contribution statement

AI systems used: OpenAI Codex (experiment execution and code) and Anthropic Claude (study design, analysis review and writing). Research questions were shared between the author and these systems; study designs were drafted mainly by AI systems under the author's direction and approved by the author before each run; execution was carried out by AI systems; analysis and writing of the individual reports were shared. This synthesis and the exploratory trace-length analysis were carried out with Anthropic Claude from the individual reports and data, at the author's direction; the trace-length question was raised by the author. The author takes responsibility for the content.

Data availability

Preregistrations, code, raw data and individual study reports are archived on Zenodo [1–4] and mirrored at `https://github.com/arsumed-del/affect-under-constraint`. The trace-length analysis plan, script and results are included with this version.

Ethics note

Steering was applied to small open-weight models at doses chosen by preregistered calibration rules. Whether interventions of this kind bear on model welfare is an open question, noted here for completeness.

Competing interests

None.

References

[1] A. Su. Affect Under Constraint: Persistence of a steered affective representation in instruction-tuned language models (Studies 1–2). Preprint, v3, Zenodo, 2026. doi:10.5281/zenodo.23182205.

[2] A. Su. Affect Under Constraint: A steered affective state carries forward through a language model's own reasoning (Study 3). Preprint, Zenodo, 2026. doi:10.5281/zenodo.23323490.

[3] A. Su. Affect Under Constraint: Conversational placement and fixed-answer tests of hostility-related carryover (Studies 4–6). Preprint, Zenodo, 2026. doi:10.5281/zenodo.23747766.

[4] A. Su. Affect Under Constraint: How warmth-related constraints change the expression of a hostility-associated influence in a language model's reasoning (Studies 7–8). Preprint, Zenodo, 2026. doi:10.5281/zenodo.23876548.

[5] M. Sun, T. Li, Y. Zheng, Z. Zhou, A. Liu, X. Liu, Y. Liu. How emotion shapes the behavior of LLMs and agents: A mechanistic study. arXiv:2604.00005, 2026.

[6] Anthropic Interpretability Team. Emotion concepts and their function in a large language model. 2 April 2026. `https://www.anthropic.com/research/emotion-concepts-function`.

[7] J. Jeong. Extracting and steering emotion representations in small language models: A methodological comparison. arXiv:2604.04064, 2026 (Model Medicine series, Paper 6).

[8] J. Hartmann. Emotion English DistilRoBERTa-base. Hugging Face model, 2022. `https://huggingface.co/j-hartmann/emotion-english-distilroberta-base`.

[9] A. Su. Computational psychodynamics: Defense mechanisms as transformation operators in cognitive-affective dynamics. Preprint, Zenodo, 2026. doi:10.5281/zenodo.21925669.

[10] A. Su. Affective transformation beyond discrete defenses: Toward a continuous transformation framework. Preprint, Zenodo, 2026. doi:10.5281/zenodo.19521936.

[11] A. Su. A minimal recursive loop model of mood and anxiety disorders. Preprint, Zenodo, 2026. doi:10.5281/zenodo.20300446.

[12] A. Su. Artificial worlds as experimental systems for psychological theory. Preprint, Zenodo, 2026. doi:10.5281/zenodo.23072283.